

Weighted Hypothesis Testing

Larry Wasserman and Kathryn Roeder¹
Carnegie Mellon University

April 7, 2006

The power of multiple testing procedures can be increased by using weighted p-values (Genovese, Roeder and Wasserman 2005). We derive the optimal weights and we show that the power is remarkably robust to misspecification of these weights. We consider two methods for choosing weights in practice. The first, external weighting, is based on prior information. The second, estimated weighting, uses the data to choose weights.

1 Introduction

The power of multiple testing procedures can be increased by using weighted p-values (Genovese, Roeder and Wasserman 2005). Dividing each p-value P by a weight w increases the probability of rejecting some hypotheses. Provided the weights have mean one, familywise error control methods and false discovery control methods maintain their frequentist error control guarantees.

The first such weighting scheme appears to be Holm (1979). Related ideas are in Benjamini and Hochberg (1997) and Chen et al (2000). There are, of course, other ways to improve power aside from weighting. Some notable recent approaches include Rubin, van der Laan and Dudoit (2005), Storey (2005), Donoho and Jin (2004) and Signoravitch (2006). Of these, our approach is closest to Rubin, van der Laan and Dudoit (2005), hereafter, RVD. In fact, the optimal weights derived here, if re-expressed as cutoffs for test statistics, turn out to be identical to the cutoffs derived in RVD. Our main contributions beyond RVD are (i) a careful study of potential power losses due to departures from the optimal weights, (ii) robustness properties of weighted methods, and (iii) recovering power after using data splitting to estimate the weights. An important distinction between this paper and RVD versus Storey (2005) is that Storey uses a slightly different loss function and he requires a common cutoff for all test statistics. This allows him to make an elegant connection with the Neyman-Pearson lemma. In particular, his method automatically adapts from one-sided testing to two-sided testing depending on the configuration of means. Signoravitch (2006) uses invariance arguments to find powerful test statistics for multiple testing when the underlying tests are multivariate.

In this paper we show that the optimal weights form a one parameter family. We also show the power is very robust to misspecification of the weights. In particular, we show that (i) sparse

¹Research supported by National Institute of Mental Health grants MH057881, MH066278, MH06329 and NSF Grant AST 0434343. The authors thank Jamie Robins for helping us to clarify several issues.

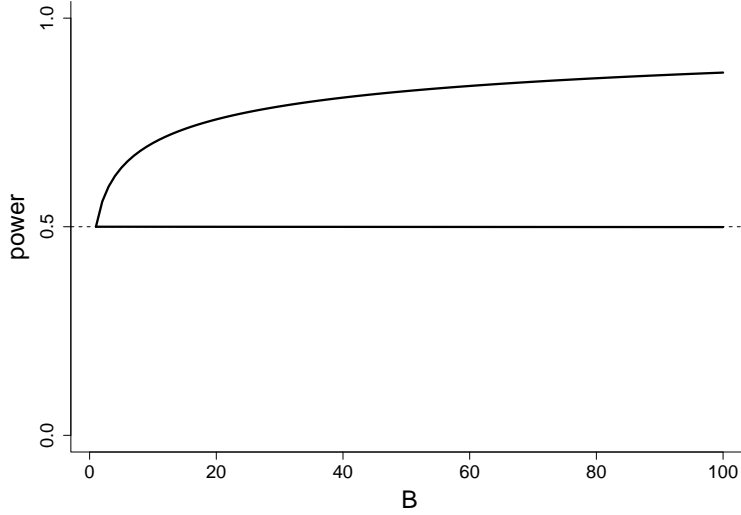


Figure 1: Power gain/loss for weighting a single hypothesis. In this example, an unweighted hypothesis has power $1/2$. The weights are $w_0 < 1 < w_1$ with $w_1/w_0 = B$. The top line shows the power when the alternative is given the correct weight w_1 . The bottom line, which is nearly indistinguishable from $1/2$, shows the power when the alternative is given the incorrect weight w_0 . As B increases, the power gain increases sharply while the power loss remains nearly constant.

weights (a few large weights and minimum weight close to 1) lead to huge power gain for well specified weights, but minute power loss for poorly specified weights; and (ii) in the non-sparse case, under weak conditions, the worst case power loss for poorly specified weights is typically better than the power using equal weights. In fact, the power is degraded at most by a factor of about $\gamma/(1 - a)$ where a is the fraction of nonnulls and γ is the fraction of nulls that are mistaken for alternatives. Figure 1 shows the sparse case. The top line shows power from correct weighting while the bottom line shows power from incorrect weighting. We see that the power gains overwhelm the potential power loss. Figure 2 shows the non-sparse case. The plots on the left show the power as a function of the alternative mean ξ . The dark solid line shows the lowest possible power assuming the weights were estimated as poorly as possible. The lighter solid line is the power of the unweighted (Bonferroni) method. The dotted line shows the power under theoretically optimal weights. The worst case weighted power is typically close to or larger than the Bonferroni power except for large ξ when they are both large.

We consider two methods for choosing the weights: (i) external weights, where prior information (based on scientific knowledge or prior data) singles out specific hypotheses and (ii) estimated weights where the data are used to construct weights. External weights are prone to bias while estimated weights are prone to variability. The two robustness properties reduce concerns about bias and variance.

An example of external weighting is the following. We have test statistics $\{T_j : j = 1, \dots, m\}$

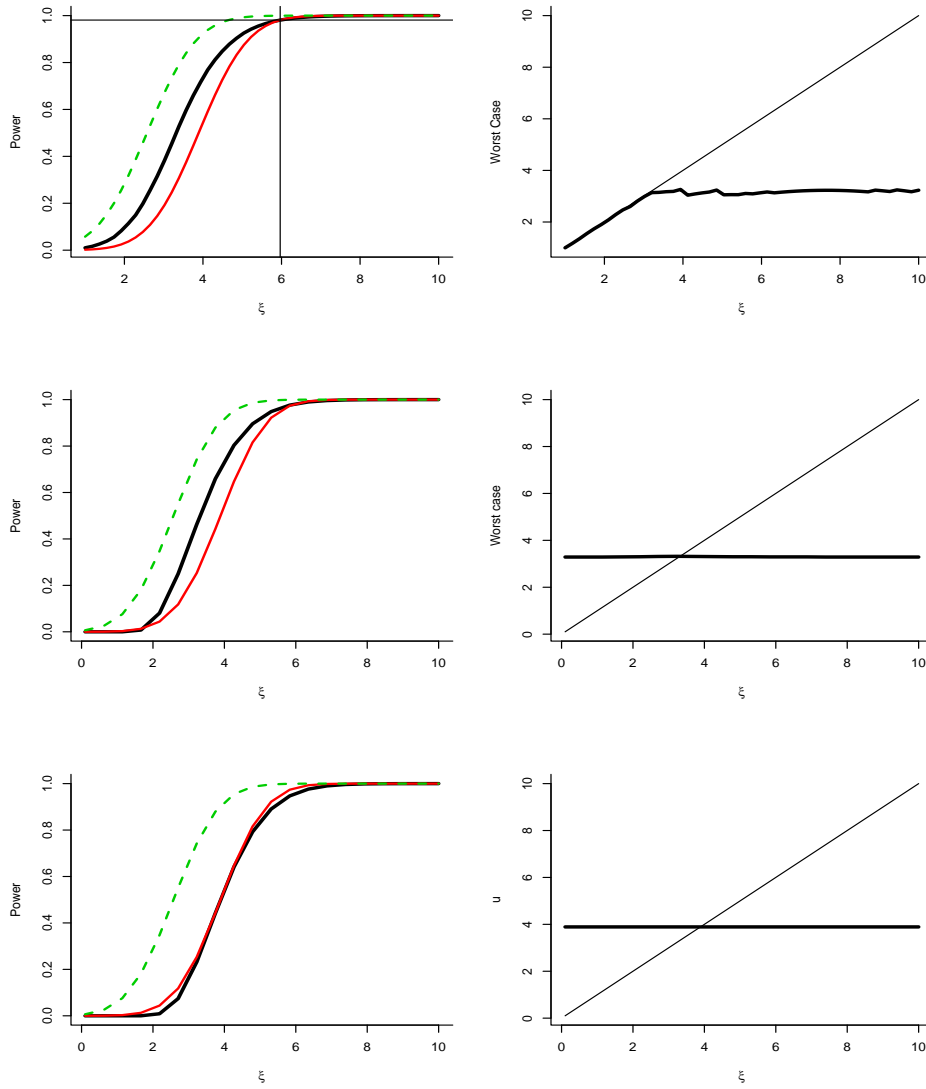


Figure 2: Power as a function of the alternative mean ξ . In these plots, $a = .01$, $m = 1000$ and $\alpha = 0.05$. There are $(1 - a)m$ nulls and ma alternatives with mean ξ . The left plots show what happens when the weights are incorrectly computed assuming that a fraction γ of nulls are actually alternatives with mean u . In the top plot, we restrict $0 < u < \xi$. In the second and third plot, no restriction is placed on u . The top and middle plot have $\gamma = .1$ while the third plot has $\gamma = 1 - a$ (all nulls misspecified as alternatives). The dark solid line shows the lowest possible power assuming the weights were estimated as poorly as possible. The lighter solid line is the power of the unweighted (Bonferroni) method. The dotted line is the power under the optimal weights. The vertical line in the top plot is at ξ_* . The weighted method beats unweighted for all $\xi < \xi_*$. The right plot shows the least favorable u as a function of ξ . That is, mistaking γm nulls for alternatives with mean u leads to the worst power. Also shown is the line $u = \xi$.

associated with spatial locations $\{s_j : j = 1, \dots, m\}$ where $s_j \in [0, L]$, say. These could be association tests for markers on a genome. The number of tests m is large, on the order of 100,000 for example. Each T_j is used to test the null hypothesis that $\theta_j = \mathbb{E}(T_j) = 0$. Prior data is in the form of a smooth stochastic process $\{Z(s) : s \in [0, L]\}$. This might be from a whole genome linkage scan. At alternatives, the mean $\mu(s) = \mathbb{E}(Z(s))$ is a large positive value; however, due to correlation, at nulls close to alternatives, $\mu(s)$ is also non-zero. Peaks in the process $Z(s)$ provide approximate information about the location of alternatives. We want to use the process Z to generate reasonable weights for the test statistics.

When external weights are not available, the optimal weights can be estimated from the data. One approach is to use data splitting (RVD) using a fraction of the data to estimate the weights and the remainder to test. For example, consider the two-stage genome-wide association study (e.g., Thomas et al. 2005) for which a sample of n subjects is split into two subsets. Using the first subset, we obtain test statistics $\{T_j : j = 1, \dots, m\}$ associated with locations $\{s_j : j = 1, \dots, m\}$. Typically only the second subset of data are used in the final analysis. Building on the ideas of Skol et al. (2006), we take the two-stage study design further, exploring how the first set of data can be utilized to formulate weights, and the full data set can be used for testing.

2 Weighted Multiple Testing

We are given hypotheses $H = (H_1, \dots, H_m)$ and standardized test statistics $T = (T_1, \dots, T_m)$ where $T_j \sim N(\xi_j, 1)$. (The methods can be extended for nonnormal test statistics but we do not consider that case here.) For a two-sided hypothesis, $H_j = 1$ if $\xi_j \neq 0$ and $H_j = 0$ otherwise. For the sake of parsimony, unless otherwise noted, results will be stated for a one-sided test where $H_j = 1$ if $\xi_j > 0$ although the results extend easily to the two-sided case. Let $\theta = (\xi_1, \dots, \xi_m)$ denote the vector of means.

The original data are often of the form

$$\mathbb{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix} \quad (1)$$

where the j^{th} test statistic T_j is based on the j^{th} column of \mathbb{X} . Usually, T_j is of the form $T_j = \sqrt{n_j} \bar{X}_j / \sigma_j$ where \bar{X}_j is approximately (or exactly) $N(\gamma_j, \sigma_j^2/n_j)$ and the noncentrality parameter is $\xi_j = \sqrt{n_j} \gamma_j / \sigma_j$.

The p-values associated with the tests are $P = (P_1, \dots, P_m)$ where $P_j = \overline{\Phi}(T_j)$, $\overline{\Phi} = 1 - \Phi$ and Φ denotes the standard Normal CDF. Let

$$P_{(1)} \leq \dots \leq P_{(m)}$$

denote the sorted p-values and let

$$T_{(1)} \geq \dots \geq T_{(m)}$$

denote the sorted test statistics.

A *rejection set* \mathcal{R} is a subset of $\{1, \dots, m\}$. Say that \mathcal{R} *controls familywise error at level α* if $\mathbb{P}(\mathcal{R} \cap \mathcal{H}_0) \leq \alpha$ where $\mathcal{H}_0 = \{j : H_j = 0\}$. The *Bonferroni rejection set* is

$$\mathcal{R} = \{j : P_j < \alpha/m\} = \{j : T_j > z_{\alpha/m}\} \quad (2)$$

where we use the notation $z_\beta = \overline{\Phi}^{-1}(\beta)$.

The weighted Bonferroni procedure of Genovese, Roeder and Wasserman (2005) is as follows. Specify nonnegative weights $w = (w_1, \dots, w_m)$ and reject hypothesis H_j if

$$j \in \mathcal{R} = \left\{ j : \frac{P_j}{w_j} \leq \frac{\alpha}{m} \right\}. \quad (3)$$

As long as $m^{-1} \sum_j w_j = 1$, the rejection set \mathcal{R} controls familywise error at level α . For completeness, we provide the proof. (All further proofs are in the appendix.)

Lemma 2.1 *If $m^{-1} \sum_j w_j = 1$, then the rejection set \mathcal{R} controls familywise error at level α .*

Proof. The familywise error is

$$\begin{aligned} \mathbb{P}((\mathcal{R} \cap \mathcal{H}_0) > 0) &= \mathbb{P}\left(P_j \leq \frac{\alpha w_j}{m} \text{ for some } j \in \mathcal{H}_0\right) \\ &\leq \sum_{j \in \mathcal{H}_0} \mathbb{P}\left(P_j \leq \frac{\alpha w_j}{m}\right) = \frac{\alpha}{m} \sum_{j \in \mathcal{H}_0} w_j \leq \alpha \overline{w} = \alpha. \quad \blacksquare \end{aligned}$$

Genovese, Roeder and Wasserman (2005) also showed that false discovery methods benefit by weighting. Recall that the false discovery proportion (FDP) is

$$\text{FDP} = \frac{\text{number of false rejections}}{\text{number of rejections}} = \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}|} \quad (4)$$

where the ratio is defined to be 0 if the denominator is 0. The false discovery rate (FDR) is $\text{FDR} = \mathbb{E}(\text{FDP})$. Benjamini and Hochberg (1995) proved $\text{FDR} \leq \alpha$ if $\mathcal{R} = \{j : P_{(j)} \leq T\}$ where $T = \max\{j : P_{(j)} \leq j\alpha/m\}$. Genovese, Roeder and Wasserman (2004) showed that $\text{FDR} \leq \alpha$ if the P_j 's are replaced by $Q_j = P_j/w_j$ as long as $m^{-1} \sum_j w_j = 1$ as before. This paper will focus only on familywise error. Similar results hold for FDR and will be in a followup paper.

3 Power and Optimality

3.1 Power

Before weighting, that is using weight 1, the power of a single, one-sided alternative is

$$\pi(\xi_j, 1) = \mathbb{P}(T_j > z_{\alpha/m}) = \overline{\Phi}(z_{\alpha/m} - \xi_j). \quad (5)$$

The power² in the weighted case is

$$\pi(\xi_j, w_j) = \mathbb{P}\left(P_j < \frac{\alpha w_j}{m}\right) = \mathbb{P}\left(T_j > \overline{\Phi}^{-1}\left(\frac{\alpha w_j}{m}\right)\right) = \overline{\Phi}\left(\overline{\Phi}^{-1}(z_{\alpha w_j/m}) - \xi_j\right). \quad (6)$$

Weighting increases the power when $w_j > 1$ and decreases the power when $w_j < 1$.

Given $\theta = (\xi_1, \dots, \xi_m)$ and $w = (w_1, \dots, w_m)$ we define the *average power*

$$\frac{1}{m} \sum_{j=1}^m \pi(\xi_j, w_j) I(\xi_j > 0). \quad (7)$$

More generally, if ξ is drawn from a distribution Q and $w = w(\xi)$ is a weight function we define the average power

$$\int \pi(\xi, w(\xi)) I(\xi > 0) dQ(\xi). \quad (8)$$

If we take Q to be the empirical distribution of (ξ_1, \dots, ξ_m) then this reduces to the previous expression. In this case we require $w(\xi) \geq 0$ and $\int w(\xi) dQ(\xi) = 1$.

3.2 Optimality and Robustness

In the following theorem we see that the set of optimal weight functions form a one parameter family indexed by a constant c .

Theorem 3.1 *Given $\theta = (\xi_1, \dots, \xi_m)$, the optimal weight vector $w = (w_1, \dots, w_m)$ that maximizes the average power subject to $w_j \geq 0$ and $m^{-1} \sum_{j=1}^m w_j = 1$ is $w = (\rho_c(\xi_1), \dots, \rho_c(\xi_m))$ where*

$$\rho_c(\xi) = \left(\frac{m}{\alpha}\right) \overline{\Phi}\left(\frac{\xi}{2} + \frac{c}{\xi}\right) I(\xi > 0), \quad (9)$$

²For a two-sided alternative the power is

$$\pi(\xi_j, w_j) = \overline{\Phi}\left(\overline{\Phi}^{-1}\left(\frac{\alpha w_j}{2m}\right) - \xi_j\right) + \overline{\Phi}\left(\overline{\Phi}^{-1}\left(\frac{\alpha w_j}{2m}\right) + \xi_j\right).$$

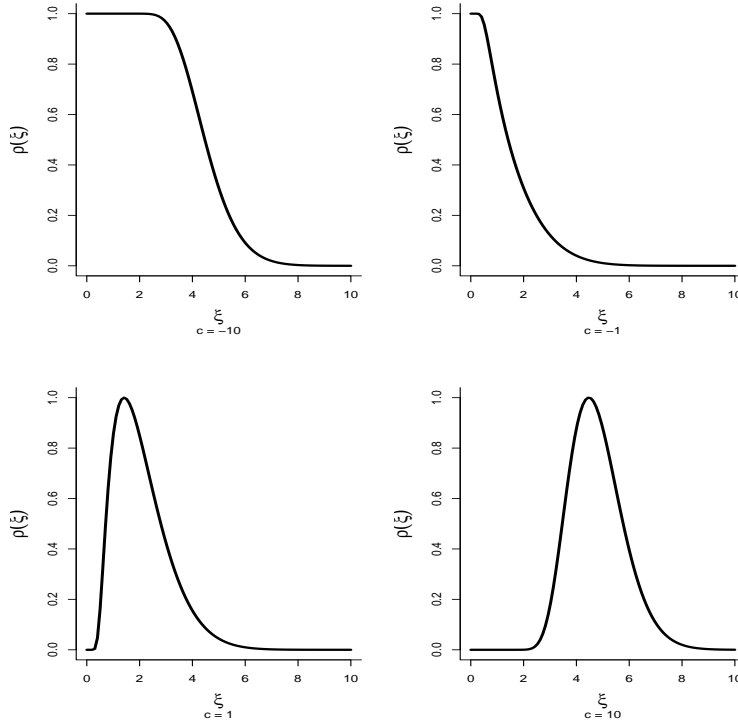


Figure 3: Optimal weight function $\rho_c(\xi)$ for various c . In each case $m = 1000$ and $\alpha = 0.05$. The functions are normalized to have maximum 1.

and $c \equiv c(\theta)$ is defined by the condition

$$\frac{1}{m} \sum_{j=1}^m \rho_c(\xi_j) = 1. \quad (10)$$

The proof is in the appendix. Some plots of the function $\rho_c(\xi)$ for various values of c are shown in Figure 3. In these plots, the function is normalized to have maximum 1 for easier visualization. The result generalizes to the case where the alternative means are random variables with distribution Q in which case c is defined by

$$\int \rho_c(\xi) dQ(\xi) = 1. \quad (11)$$

Remark. Rejecting when $P_j/w_j \leq \alpha/m$ is the same as rejection when $Z_j > \xi_j/2 + c/\xi_j$. This is identical to the result of Rubin, van der Laan and Dudoit (2005), obtained independently. The remainder of the paper, which shows some good properties of the weighted method, can thus also be considered as providing support for their method. In particular, they noted in their simulations then even poorly specified estimates of the cutoffs $\xi_j/2 + c/\xi_j$ can still perform well. This paper provides insight into why that is true.

From (6) and (9) we have immediately:

Lemma 3.2 *The power at an alternative with mean ξ under optimal weights is $\bar{\Phi}(c/\xi - \xi/2)$. The average power under optimal weights, which we call the **oracle power**, is*

$$\frac{1}{m} \sum_{j=1}^m \bar{\Phi} \left(\frac{c}{\xi_j} - \frac{\xi_j}{2} \right) I(\xi_j > 0). \quad (12)$$

The oracle power is not attainable since the optimal weights depend on $\theta = (\xi_1, \dots, \xi_m)$ or, equivalently, on Q . In practice, the weights will either be chosen by prior information or by estimating the ξ 's. This raises the following question: how sensitive is the power to correct specification of the weights? Now we show that the power is very robust to weight misspecification.

The weights themselves can be very sensitive to changes in θ . Consider the following example. Suppose that $\theta = (\xi_1, \dots, \xi_m)$ where each ξ is equal to either 0 or some fixed number ξ . The empirical distribution of the ξ_j 's is thus $Q = (1 - a)\delta_0 + a\delta_\xi$ where δ denotes a point mass and a is the fraction of nonzero means. The optimal weights are 0 for $\xi_j = 0$ and $1/a$ for $\xi_j = \xi$. Let $\tilde{Q} = (1 - a - \gamma)\delta_0 + \gamma\delta_u + a\delta_\xi$ where u is a small positive number. Since we have only moved the mass at 0 to u , and u is small, we would hope that $w(\xi)$ will not change much. But this is not the case. Set

$$\xi = A + \sqrt{A^2 - 2c}, \quad u = B - \sqrt{B^2 - 2c} \quad (13)$$

where

$$A = \bar{\Phi}^{-1} \left(\frac{\alpha}{m(\gamma K + a)} \right), \quad B = \bar{\Phi}^{-1} \left(\frac{K\alpha}{m(\gamma K + a)} \right), \quad (14)$$

yields weights w_0 and w_1 on u and ξ such that $w_0/w_1 = K$. For example, take $m = 1000$, $\alpha = 0.05$, $a = .1$, $\gamma = .1$, $K = 1000$, and $c = .1$. Then $u = .03$ and $\xi = 9.8$. The optimal weight on ξ under Q is 10 but under \tilde{Q} it is .00999 and so is reduced by a factor of 1001. More generally we have the following result which shows that the weights are, in a certain sense, a discontinuous function of θ .

Lemma 3.3 *Fix α and m . For any $\delta > 0$ and $\epsilon > 0$ there exists $Q = (1 - a)\delta_0 + a\delta_\xi$ and $\tilde{Q} = (1 - a - \gamma)\delta_0 + \gamma\delta_u + a\delta_\xi$ such that*

$$d(Q, \tilde{Q}) < \delta, \quad \text{and} \quad \frac{\tilde{\rho}(\xi)}{\rho(\xi)} < \epsilon \quad (15)$$

where $a = \alpha/4$, $d(Q, \tilde{Q}) = \sup_\xi |Q(-\infty, \xi] - \tilde{Q}(-\infty, \xi]|$ is the Kolmogorov-Smirnov distance, ρ is the optimal weight function for Q and $\tilde{\rho}$ is the optimal weight function for \tilde{Q} .

Fortunately, this problem is not serious since it is possible to have high power even with poor weights. In fact, the power of the weighted method has the following two robustness properties:

Property I: Sparse weights (minimum weight close to 1) are highly robust. If most weights are less than 1 and the minimum weight is close to 1 then correct specification (large weights on alternatives) leads to large power gains but incorrect specification (large weights on nulls) leads to little power loss.

Property II: Worst case analysis. Weighted hypothesis testing, even with poorly chosen weights, typically does as well or better than Bonferroni except when the the alternative means are large, in which both have high power.

Let us now make the these statements precise. Also, see Genovese, Roeder and Wasserman (2006) and Roeder, Bacanu, Wasserman and Devlin (2006) for other results on the effect of weight misspecification.

Property I. Consider first the case where the weights take two distinct values and the alternatives have a common mean ξ . Let ϵ denote the fraction of hypotheses given the larger of the two values of the weights B . Then, the weight vector w is proportional to

$$\underbrace{(B, \dots, B)}_{k \text{ terms}}, \underbrace{(1, \dots, 1)}_{m-k \text{ terms}}$$

where $k = \epsilon m$ and $B > 1$ and hence the normalized weights are

$$w = \underbrace{(w_1, \dots, w_1)}_{k \text{ terms}}, \underbrace{(w_0, \dots, w_0)}_{m-k \text{ terms}}$$

where

$$w_1 = \frac{B}{\epsilon B + (1 - \epsilon)}, \quad w_0 = \frac{1}{\epsilon B + (1 - \epsilon)}.$$

We say that the weights are sparse if ϵ is small, that is, if most weights are near 1.

Consider an alternative with mean ξ . The power gain by correct weighting is the power under weight w_1 minus the unweighted power $\pi(\xi, w_1) - \pi(\xi, 1)$. Similarly, the power loss for incorrect weighting is $\pi(\xi, 1) - \pi(\xi, w_0)$. The gain minus the loss, which we call the robustness function, is

$$R(B, \epsilon) \equiv \left(\pi(\xi, w_1) - \pi(\xi, 1) \right) + \left(\pi(\xi, 1) - \pi(\xi, w_0) \right) \quad (16)$$

$$= \bar{\Phi}(z_{\alpha w_1/m} - \xi) + \bar{\Phi}(z_{\alpha w_0/m} - \xi) - 2\bar{\Phi}(z_{\alpha/m} - \xi). \quad (17)$$

The gain outweighs the loss if and only if $R(B, \epsilon) > 0$. This is illustrated in figures 1 and 4.

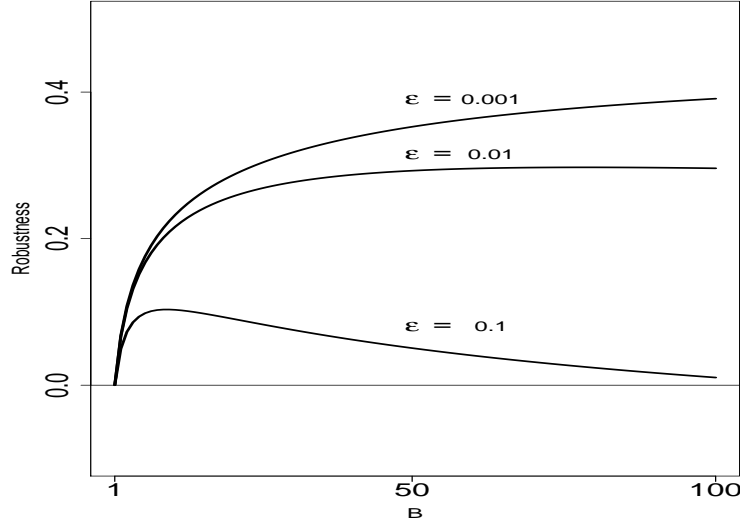


Figure 4: Robustness function for $m = 1000$. In this example, $\xi = z_{\alpha/m}$ which has power 1/2 without weighting. The gain of correct weighting far outweighs the loss for incorrect weighting as long as the fraction of large weights ϵ is small.

Theorem 3.4 Fix $B > 1$. Then, $\lim_{\epsilon \rightarrow 0} R(B, \epsilon) > 0$. Moreover, there exists $\epsilon^*(B) > 0$ such that $R(B, \epsilon) > 0$ for all $\epsilon < \epsilon^*(B)$.

We can generalize this beyond the two-valued case as follows. Let w be any weight vector such that $m^{-1} \sum_j w_j = 1$. Now define the (worst case) robustness function

$$R(\xi) \equiv \min_{\{j: w_j > 1, H_j=1\}} \{\pi(\xi, w_j) - \pi(\xi, 1)\} - \max_{\{j: w_j < 1, H_j=1\}} \{\pi(\xi, 1) - \pi(\xi, w_j)\}. \quad (18)$$

We will see that $R(\xi) > 0$ under weak conditions and that the maximal robustness is obtained for ξ near the Bonferroni cutoff $z_{\alpha/m}$.

Theorem 3.5 A necessary and sufficient condition for $R(\xi) > 0$ is

$$R_{b,B}(\xi) \equiv \Phi\left(z_{\alpha B/m} - \xi\right) + \Phi\left(z_{\alpha b/m} - \xi\right) - 2\Phi\left(z_{\alpha/m} - \xi\right) \leq 0 \quad (19)$$

where $B = \min\{w_j : w_j > 1\}$, $b = \min\{w_j\}$. Moreover,

$$R_{b,B}(\xi) = -\Delta(\xi) + O(1 - b) \quad (20)$$

where

$$\Delta(\xi) = \left(\Phi\left(z_{\alpha/m} - \xi\right) - \Phi\left(z_{\alpha B/m} - \xi\right) \right) > 0 \quad (21)$$

and, as $b \rightarrow 1$, $\mu(\{\xi : R(\xi) < 0\}) \rightarrow 0$ and $\inf_{\xi} R(\xi) \rightarrow 0$.

The theorem is illustrated in Figure 5. We see that there is overwhelming robustness as long as the minimum weight is near 1. Even in the extreme case $b = 0$, there is still a safe zone, an interval of values of ξ over which $R(\xi) > 0$.

Lemma 3.6 *Suppose that $B \geq 2$. Then there exists $\xi_* > 0$ such that $R_{B,b}(\xi) > 0$ for all $0 \leq \xi \leq \xi_*$ and all b . An upper bound on ξ_* is $z_{\alpha/m} - 1/(z_{\alpha/m} - z_{B\alpha/m})$.*

Property II. Even if the weights are not sparse, the power of the weighted test cannot be too bad as we now show. To begin, assume that each mean is either equal to 0 or ξ for some fixed $\xi > 0$. Thus, the empirical distribution is

$$Q = (1 - a)\delta_0 + a\delta_\xi \quad (22)$$

where δ denotes a point mass and a is the fraction of nonzero ξ_j 's. The optimal weights are $1/a$ for hypotheses whose mean is ξ . To study the effect of misspecification error, consider the case where $b = \gamma m$ nulls are mistaken for alternatives with mean $u > 0$. This corresponds to misspecifying Q to be

$$\tilde{Q} = (1 - a - \gamma)\delta_0 + \gamma\delta_u + a\delta_\xi. \quad (23)$$

We will study the effect of varying u so let $\pi(u)$ denote the power at the true alternative ξ as a function of u . Also, let π_{Bonf} denote the power using equal weights (Bonferroni). Note that changing $Q = (1 - a)\delta_0 + a\delta_\xi$ to $Q = (1 - a)\delta_0 + a\delta_{\xi'}$ for $\xi' \neq \xi$ does not change the weights.

As the weights are a function of c , we first need to find c as a function of u . The normalization condition (10) reduces to

$$\gamma\bar{\Phi}\left(\frac{u}{2} + \frac{c}{u}\right) + a\bar{\Phi}\left(\frac{\xi}{2} + \frac{c}{\xi}\right) = \frac{\alpha}{m} \quad (24)$$

which implicitly defines the function $c(u)$. First we consider what happens when u is restricted to be less than ξ .

Theorem 3.7 *Assume that $\alpha/m \leq \gamma + a \leq 1$. Let $Q = (1 - a)\delta_0 + a\delta_\xi$ and $\tilde{Q} = (1 - a - \gamma)\delta_0 + \gamma\delta_u + a\delta_\xi$ with $0 \leq u \leq \xi$. Let $C(\xi) = \sup_{0 \leq u \leq \xi} c(u)$ and define $\xi_0 = z_{\alpha/(m(\gamma+a))}$,*

1. *For $\xi \leq \xi_0$,*

$$C(\xi) = \xi\xi_0 - \xi^2/2. \quad (25)$$

For $\xi > \xi_0$, $C(\xi)$ is the solution to

$$\gamma\bar{\Phi}(\sqrt{2c}) + a\bar{\Phi}\left(\frac{c}{\xi} + \frac{\xi}{2}\right) = \frac{\alpha}{m}. \quad (26)$$

In this case, $C(\xi) = z_{\alpha/(m\gamma)}^2/2 + O(a)$.

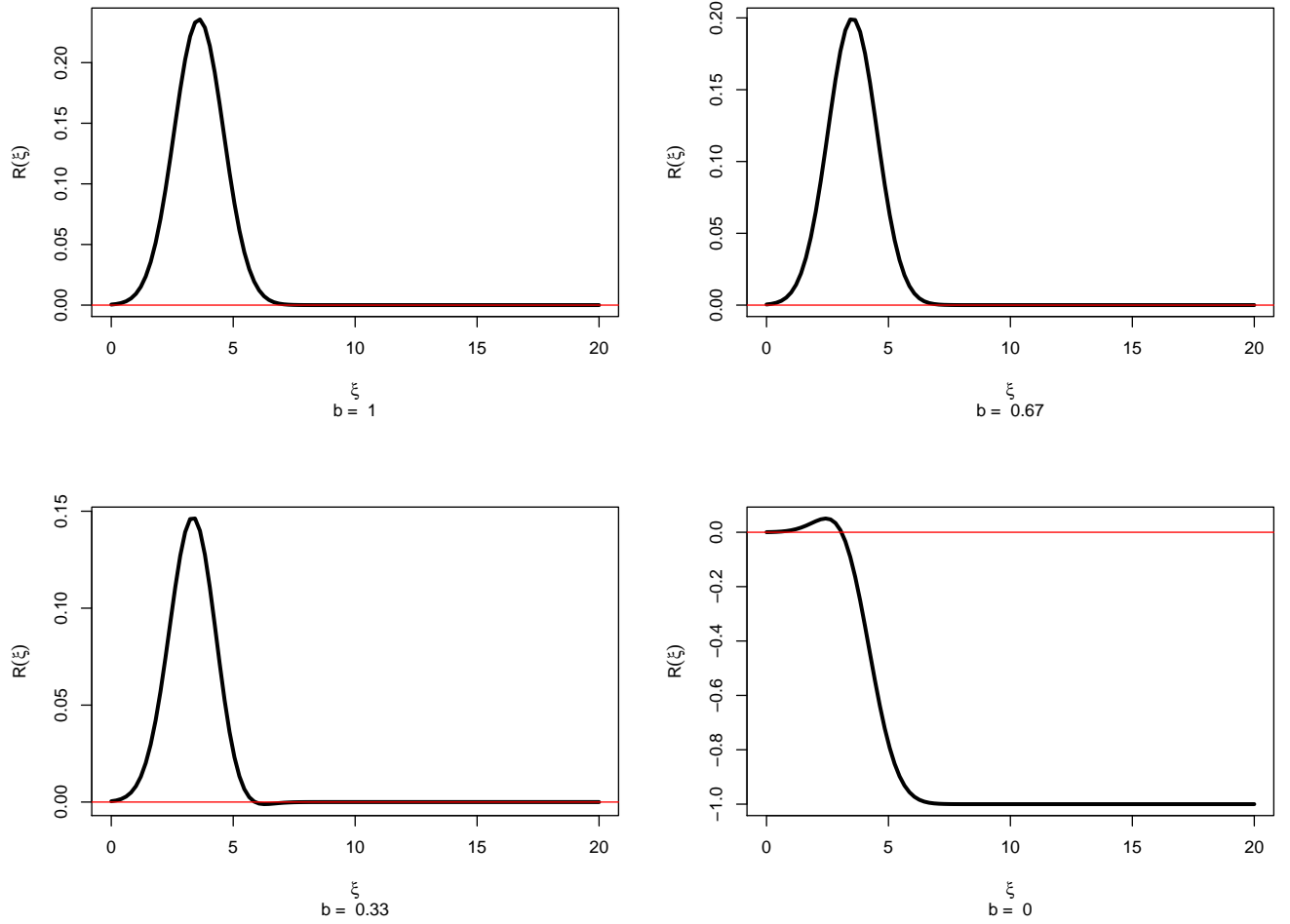


Figure 5: The robustness function $R(\xi)$ for several values of $b = \min_j w_j$. In each case, $m = 1000$, $\alpha = 0.05$, $B = 10$. Whenever $R(\xi) > 0$, power gain outweighs power loss. When b is near 1, $R(\xi) > 0$ for most ξ . Even when $b = 0$ there is a safe zone including $\xi = 0$ as long as $B \geq 2$.

2. Let

$$\xi_* = z_{\alpha/m} + \sqrt{z_{\alpha/m}^2 - z_q^2}, \quad \text{where } q = \frac{\alpha(1-a)}{m\gamma}. \quad (27)$$

For $\xi < \xi_*$,

$$\inf_{0 < u < \xi} \pi(u) \geq \pi_{\text{Bonf}}. \quad (28)$$

For $\xi \geq \xi_*$ we have

$$\inf_{0 < u < \xi} \pi(u) \geq \bar{\Phi} \left(\frac{z_{\alpha/(m\gamma)}^2 - \xi_*^2}{2\xi_*} \right) - O(a) \quad (29)$$

$$\approx 1 - \bar{\Phi} \left(\sqrt{2 \log \frac{1-a}{\gamma}} \right) - O(a) \quad (30)$$

$$\geq 1 - \frac{\gamma}{1-a} - O(a). \quad (31)$$

The factor $\bar{\Phi} \left(\sqrt{2 \log \frac{1-a}{\gamma}} \right) \approx \gamma/(1-a)$ is the worst case power deficit due to misspecification. Now we drop the assumption that $u \leq \xi$.

Theorem 3.8 Let $Q = (1-a)\delta_0 + a\delta_\xi$ and let $Q_u \equiv (1-a-\gamma)\delta_0 + \gamma\delta_u + a\delta_\xi$. Let π_u denote the power at ξ using the weights computed under Q_u .

1. The least favorable u is

$$u_* \equiv \operatorname{argmin}_{u \geq 0} \pi_u = \sqrt{2c_*} = z_{\alpha/(m\gamma)} + O(a) \quad (32)$$

where c_* solves

$$\gamma \bar{\Phi}(\sqrt{2c_*}) + a \bar{\Phi} \left(\frac{\xi}{2} + \frac{c_*}{\xi} \right) = \frac{\alpha}{m} \quad (33)$$

and $c_* = z_{\alpha/(m\gamma)}^2/2 + O(a)$.

2. The minimal power is

$$\inf_u \pi_u = \bar{\Phi} \left(\frac{c_*}{\xi} - \frac{\xi}{2} \right) = \bar{\Phi} \left(\frac{z_{\alpha/(m\gamma)}^2 - \xi^2}{2\xi} \right) + O(a). \quad (34)$$

3. A sufficient condition for $\inf_u \pi_u$ to be larger than the power of the Bonferroni method is

$$\xi \geq z_{\alpha/m} + \sqrt{z_{\alpha/m}^2 - z_{\alpha/(m\gamma)}^2} + O(a). \quad (35)$$

4 Choosing External Weights

In choosing external weights, we will focus here on the two-valued case. Thus,

$$w = (\underbrace{w_1, \dots, w_1}_{k \text{ terms}}, \underbrace{w_0, \dots, w_0}_{m-k \text{ terms}}) \quad (36)$$

where $k = \epsilon m$, $w_1 = B/(\epsilon B + (1 - \epsilon))$ and $w_0 = 1/(\epsilon B + (1 - \epsilon))$. In practice, we would typically have a fixed fraction of hypotheses ϵ that we want to give more weight to. The question is how to choose B . We will focus on choosing B to produce weights with good properties at interesting values of ξ . Now large values of ξ already have high power. Very small values of ξ have extremely low power and benefit little by weighting. This leads us to focus on constructing weights that are useful for a *marginal effect*, defined as the alternative ξ_m that has power 1/2 when given weight 1. Thus, the marginal effect is $\xi_m = z_{\alpha/m}$. In the rest of this section then we assume that all nonzero ξ_j 's are equal to ξ_m . Of course, the validity of the procedure does not depend on this assumption being true.

Fix $0 < \epsilon < 1$ and vary B . As we increase B , we will eventually reach a point $B_0(\epsilon)$ where $R(B, \epsilon) < 0$ which we call turnaround point. Formally,

$$B_0(\epsilon) = \sup \left\{ B : R(B, \epsilon) > 0 \right\}. \quad (37)$$

The top panel in Figure 6 shows $B_0(\epsilon)$ versus ϵ which shows that for small ϵ we can choose B large without loss of power. The bottom panel shows $R(B, \epsilon)$ for $\epsilon = 0.1$. We suggest using $B = B_*(\epsilon)$, the value of B that maximizes $R(B, \epsilon)$.

Theorem 4.1 Fix $0 < \epsilon < 1$. As a function of B , $R(B, \epsilon)$ is unimodal and satisfies $R(1, \epsilon) = 1$, $R'(1, \epsilon) > 0$ and $R(\infty, \epsilon) < 0$. Hence, $B_0(\epsilon)$ exists and is unique. Also, $R(B, \epsilon)$ has a unique maximum at some point $B^*(\epsilon)$ and $R(B^*(\epsilon), \epsilon) > 0$.

When ϵ is very small, we can essentially choose B as large as we like. For example, suppose we want to increase the chance of rejecting one particular hypothesis so that $\epsilon = 1/m$. Then,

$$w_1 = \frac{mB}{B + m - 1} \approx B, \quad w_0 = \frac{1}{B + m - 1} \approx 1$$

and

$$\lim_{m \rightarrow \infty} \lim_{B \rightarrow \infty} \pi(\xi_j, w_1) = 1, \quad \text{while} \quad \lim_{m \rightarrow \infty} \lim_{B \rightarrow \infty} \pi(\xi_j, w_0) = \frac{1}{2}.$$

See Figure 1.

The next results show that binary weighting schemes are optimal in a certain sense. Suppose we want to have at least a fraction ϵ with high power $1 - \beta$ and otherwise we want to maximize the minimum power.

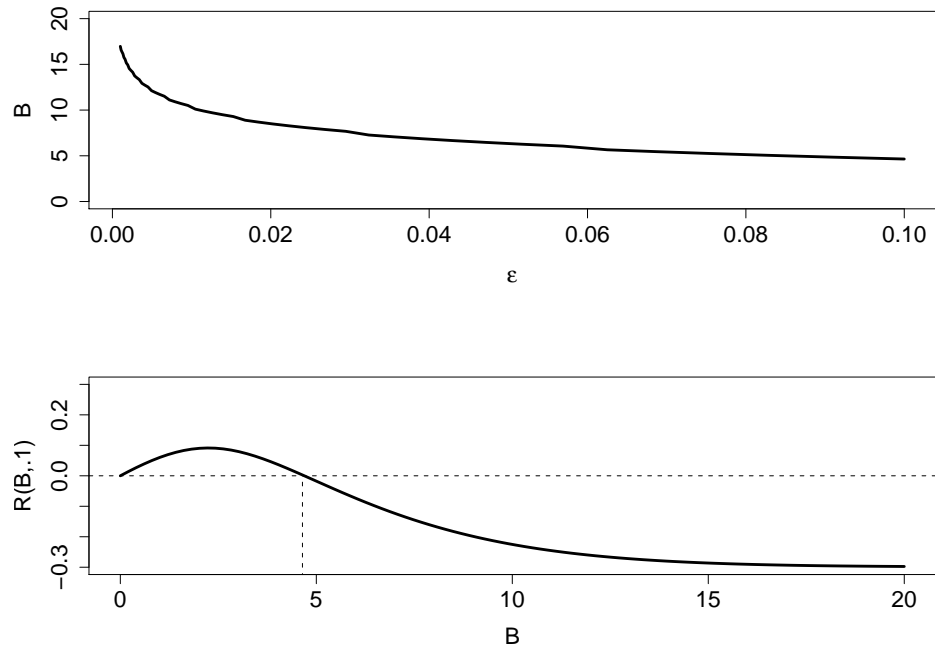


Figure 6: Top plot: $B_0(\epsilon)$ versus ϵ . Bottom plot shows $R(B, .1)$ versus B . The turnaround point $B^*(\epsilon)$ is shown with a vertical dotted line.

Theorem 4.2 Consider the following optimization problem: Given $0 < \epsilon < 1$ and $0 < \beta < 1/2$, find a vector $w = (w_1, \dots, w_m)$ that maximizes

$$\min_j \pi(\xi_m, w_j)$$

subject to

$$\bar{w} = 1, \quad \text{and} \quad \frac{\#\{j : \pi(w_j, \xi_m) \geq 1 - \beta\}}{m} \geq \epsilon.$$

The solution is given by

$$w = (\underbrace{w_1, \dots, w_1}_{k \text{ terms}}, \underbrace{w_0, \dots, w_0}_{m-k \text{ terms}}) \quad (38)$$

where $w_1 = B/(\epsilon B + (1 - \epsilon))$, $w_0 = 1/(\epsilon B + (1 - \epsilon))$, $k = \epsilon m$, $B = cm(1 - \epsilon)/(\alpha - \epsilon cm)$ and $c = \bar{\Phi}(z_{\alpha/m} + z_{1-\beta})$.

If our goal is to maximize the number of alternatives with high power while maintaining a minimum power loss, the solution is given as follows.

Theorem 4.3 Consider the following optimization problem: Given $0 < \beta < 1/2$, find a vector $w = (w_1, \dots, w_m)$ that maximizes

$$\#\{j : \pi(w_j, \xi_m) \geq 1 - \beta\} \quad (39)$$

subject to

$$\bar{w} = 1, \quad \text{and} \quad \min_j \pi(w_j, \xi_m) \geq \delta. \quad (40)$$

The solution is

$$w = (\underbrace{w_1, \dots, w_1}_{k \text{ terms}}, \underbrace{w_0, \dots, w_0}_{m-k \text{ terms}}) \quad (41)$$

where

$$w_1 = \frac{m}{\alpha} \bar{\Phi}(z_{\alpha/m} + z_{1-\beta}), \quad w_0 = \frac{m}{\alpha} \bar{\Phi}(z_{\alpha/m} + z_{\delta}), \quad \epsilon = \frac{1 - w_0}{w_1 - w_0} \quad (42)$$

and $k = m\epsilon$.

A special case that falls under this Theorem permits the minimum power to be 0. In this case $w_0 = 0$ and $\epsilon = 1/w_1$.

5 Estimated Weights

In this section we explain how to use the data to estimate the weights. There are two issues: we must ensure that the error is still controlled and avoid incurring large losses of power due to replacing θ with an estimator $\hat{\theta}$.

5.1 Validity With Estimated Weights

Data Splitting. The approach, taken by RVD, for ensuring that the error control is preserved relies on data splitting. This approach relies on normalized test statistics $T^{(l)}, T^{(2)}$ based on a partition of the data into subsets $\mathbb{X}^{(1)}, \mathbb{X}^{(2)}$ which include fractions b and $(1 - b)$ of \mathbb{X} , respectively. Note that $T_j = b^{1/2}T_j^{(1)} + (1 - b)^{1/2}T_j^{(2)}, j = 1, \dots, m$. The training data $\mathbb{X}^{(1)}$ is used to estimate the noncentrality parameter of the standardized statistic $T_j^{(1)}$, where $E[T_j^{(1)}] = \sqrt{b}\xi_j \equiv \xi_j^{(1)}$. Testing is conducted using the remaining fraction of the data $\mathbb{X}^{(2)}$. Consequently $\hat{\xi}_j^{(1)}$ must be rescaled by $r_s = (1 - b)/b$ to estimate the noncentrality parameter of the standardized statistic $T_j^{(2)}$, i.e., $\hat{\xi}_j = r_s \hat{\xi}_j^{(1)}$. The estimated weights are $\hat{w}_j(T^{(1)}) = \rho_c(\hat{\xi}_j)$. Because of the independence between the two portions of the data, familywise error is controlled at the nominal level.

Lemma 5.1 *The procedure that rejects when $P(T_j^{(2)}) < w(T^{(1)})\alpha/m$ controls the familywise error at level α .*

Recovering Power. As noted by Skol et al. (2005), data splitting incurs a loss of power because the p-values are computed using only a fraction the data. To recover this lost power, we need to use all the data to compute the p-values. When using this approach $\hat{\xi}_j^{(1)}$ must be rescaled by $r_f = b^{-1/2}$ to estimate the noncentrality parameter of the standardized statistic T_j , i.e., $\hat{\xi}_j = r_f \hat{\xi}_j^{(1)}$. As in the data splitting procedure, the estimated weights $\hat{w}_j(T^{(1)}) = \rho_c(\hat{\xi}_j)$ depend only on $\mathbb{X}^{(1)}$. To preserve error control we proceed as follows.

Theorem 5.2 *Assume $T_j^{(k)} \sim N(0, 1)$ independently for $k = 1, 2$. Suppose that weight $w(T^{(1)})$ depends only on $\mathbb{X}^{(1)}$ but the p-value $P(T_j)$ is allowed to depend on the full data \mathbb{X} . Define $c(T^{(1)})$ to solve*

$$\frac{1}{m} \sum_{j=1}^m \left(\Phi \left(\frac{\hat{\xi}_j + \frac{c(T^{(1)})}{\hat{\xi}_j} - \sqrt{b} T_j^{(1)}}{\sqrt{1 - b}} \right) \right) = \frac{\alpha}{m}. \quad (43)$$

Then the procedure that rejects when $P(T_j) < w_j(T^{(1)})\alpha/m$, where

$$w_j(T^{(1)}) = \frac{m}{\alpha} \Phi \left(\frac{\hat{\xi}_j}{2} + \frac{c(T^{(1)})}{\hat{\xi}_j} \right) \quad (44)$$

controls the familywise error at level α .

5.2 Simulations

We simulate a study with $m = 1000$ tests, yielding data of the form given in (1). A test of the hypothesis $H_0 : \xi_j \neq 0$ is performed for each j using T_j , which we assume is (approximately)

normally distributed, or equivalently $T_j^2 \sim \chi_1^2$. In our simulations we generate 50 of the 1000 tests under the alternative hypothesis with shift parameter $\xi_j = 2, 3, 4$ or 5 . We compare the power for various levels of a threshold parameter $\lambda \in (0, .5, 1, 1.5, 2, 2.5)$. We use a fraction $b = 0.5$ of the data to construct the weights and we compare four methods for estimating the noncentrality parameter:

1. The normalized statistic $\widehat{\xi}_j^{(1)} = T_j^{(1)}$.

2. Hard thresholding:

$$\widehat{\xi}_j^{(1)} = T_j^{(1)} I(|T_j^{(1)}| > \lambda). \quad (45)$$

3. Soft thresholding:

$$\widehat{\xi}_j^{(1)} = \text{sign}(T_j^{(1)}) (|T_j^{(1)}| - \lambda)_+. \quad (46)$$

4. The James-Stein estimator

$$\widehat{\xi}_j^{(1)} = \left(1 - \frac{m-2}{\sum_i (T_i^{(1)})^2} \right)_+ T_j^{(1)}. \quad (47)$$

To compute the weights we rescale $\widehat{\xi}_j^{(1)}$ by r_s or r_f as appropriate to the followup testing strategy.

Power results are displayed in Fig. 7. We first consider the power of the RVD procedure which uses the data splitting strategy and $\lambda = 0$ (Fig. 7, labeled “P” at the origin). Although RVD suggest using $\lambda = 0$, we also examine the power of this procedure for a range of values of λ . This extended RVD procedure is applying hard-thresholding to estimate θ . Next we consider the power of four testing strategies that use the full data T_j for testing rather than data splitting. The first approach (B) uses binary weights equal to m/M where $M = \sum_i I\{|T_j^{(1)}| > \lambda\}$. In this setting, when $\lambda = 0$ the method reduces to the simple one-stage Bonferroni approach. For $\lambda > 0$ it is the method of Skol et al. (2006). The remaining three approaches rely on weights estimated using hard-thresholding (H), soft-thresholding (S), or James-Stein (J). For $\lambda = 0$ methods H and S reduce to the normed sample mean which is the RVD approach adapted to incorporate the full data in the p-value. Clearly, this adaptation of the RVD method leads to a valuable increase in power. For $\lambda > 0$ this method imposes a hard threshold shrinkage effect on the parameter estimates. Notice that for any fixed value of λ , method H gives the best power. In particular, the difference in power between methods B and H illustrates the advantage of using variable weights estimated from a fraction of the data. Method H and to a lesser extent method B are nearly invariant to λ for

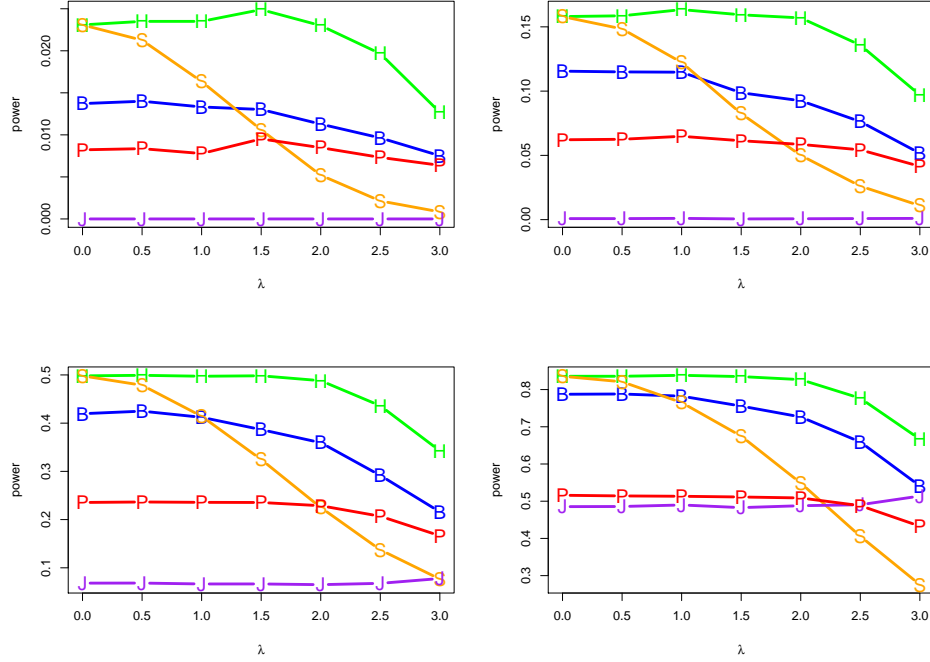


Figure 7: Power of weighted tests. From top left clockwise: $\xi = 2, 3, 4, 5$. Methods compared use weights based on hard thresholding (H), soft thresholding (S), binary weights (B), and James-Stein (J).

moderate values of the threshold parameter. In contrast, method S, relying on soft-thresholding, experiences a sharp decline in power as λ increases. Finally, the James-Stein approach clearly fails in this setting, presumably because most tests follow the null hypothesis and hence the true signals are shrunk toward 0 which diminishes the power of the procedure.

For each condition investigated the tests had size less than 0.05 as expected from the theory. The James-Stein method was most conservative.

From this experiment it appears that shrinkage only enhances power when the signal is very weak. A more careful analysis reveals that the effect of shrinkage for stronger signals is more subtle. As $\xi \rightarrow 0$, $\rho_c(\xi) \rightarrow 0$. Figure 8 shows $\rho_c(\xi)$ is close to zero for a broad range of values. Consequently, for $\lambda \leq 1.5$, the weight function performs almost the same role as the threshold parameter. Using hard-thresholding for $\lambda < 1.5$ is essentially equivalent to using using no threshold because a moderate level of shrinkage is automatically imposed by the weight function. Figure 8 also illustrates how the optimal weights vary with the signal strength (top panel has greater signal than bottom panel). Both panels indicate that larger weights are placed in the midrange of signal

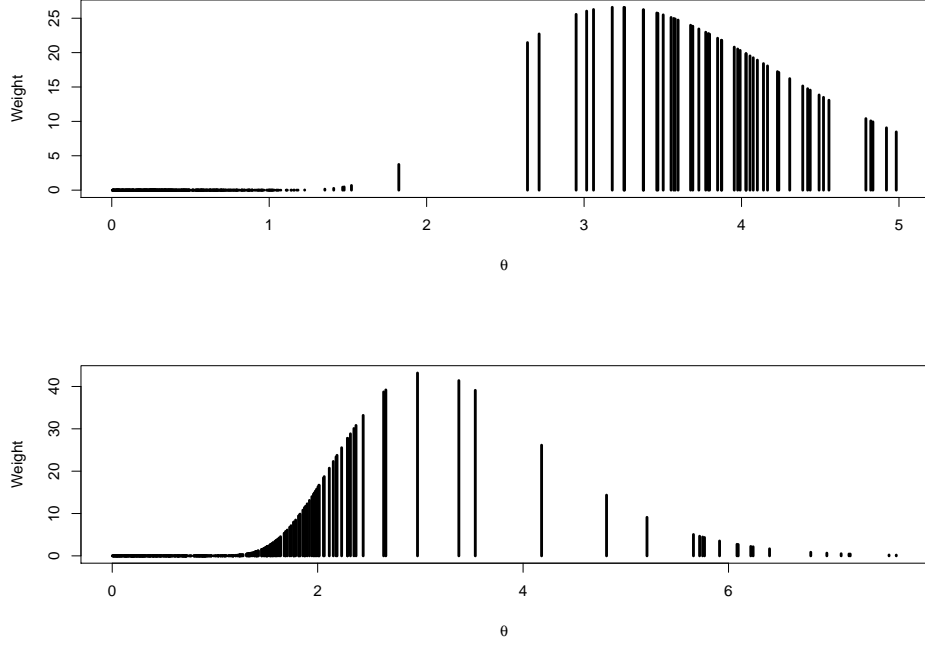


Figure 8: Distribution of weights for two sets of data.

strength. Essentially no weight is wasted on tests with small signals ($\xi < 1.5$) because these tests are not likely to yield significant results. The bottom panel shows that large weights are also not wasted on signals so strong that the tests can easily be rejected even without up-weighting ($\xi > 6$). The top panel places its largest weights between 2.5 and 4. The bottom panel has fewer signals in this range and hence stronger weights can be applied to signals between 2 and 2.5. Both panels indicate near 0 weights would be applied to tests with signals near 0.

6 Discussion

An interesting connection can be made between weights based on threshold-estimators and two-stage experimental designs that perform only a subset of the tests in stage two, based on the results obtained from stage one. The simplest example of this type of two-stage testing is the two-stage Bonferroni procedure, for which the training data $\mathbb{X}^{(1)}$ is used to determine the M elements in $\Lambda = \{j : |T_j^{(1)}| > \lambda\}$; $\mathbb{X}^{(2)}$ is only measured for these columns. A Bonferroni correction with $\alpha/(2M)$ controls FWER at level α for two-sided testing in this setting. In essence this approach is a weighted test with weights equal to m/M for the elements in Λ and zero else where.

While the classic two-stage approach uses $\mathbb{X}^{(1)}$ for training, and $\mathbb{X}^{(2)}$ for testing, an alternative is to use the training data to determine the weights and then use all of the data to conduct the tests. This strategy was recently investigated by Skol et al. (2006), using constant weights. These authors use the training data to determine Λ and then apply weights equal to m/M to the M tests determined in stage one. This full data approach proved to be considerably more powerful than the two-stage Bonferroni approach in simulations.

For hard and soft-thresholding, $\hat{\xi}_j = 0$ for any $|T_j^{(1)}| < \lambda$. From (9) it follows that the weights for any test with $\hat{\xi}_j = 0$ are 0 and the rejection region is $Z_0 = \infty$. Hence, a procedure using $w_j = 0$ for columns with $\hat{\xi}_j = 0$ is equivalent to a truncation procedure that tests only those columns in Λ . In practice, λ can be chosen to optimize power or to constrain the experimental budget. It is worth noting that in some experimental settings, such as those described by Skol et al., this experimental design can lead to considerable savings of effort and resources. Our results suggest that this savings can be gleaned without losing measurable power.

The same ideas used here can be applied to other testing methods to improve power. In particular, weights can be added to the FDR method, Holm's stepdown test, and the Donoho-Jin (2004) method. Weighting ideas can also be used for confidence intervals. We plan to present the details for the other methods in a followup paper. Another item to be addressed in future work is the connection with Bayesian methods.

As we noted, using weights is equivalent to using a separate rejection cutoff for each statistic. The methods of Storey (2005) and Signoravich (2006) find optimal cutoffs when the cutoffs are constrained. There is undoubtedly a bias-variance tradeoff. These constrained methods can estimate optimal cutoffs well (low variance) but they will not achieve the oracle power obtained here since they are by design biased away from these separate cutoffs. Future work should be directed at comparing these approaches and developing methods that lie in between these extremes.

7 Appendix

Proof of Theorem 3.1. Let A denote the set of hypotheses with $\xi_j > 0$. Power is optimized if $w_j = 0$ for $j \notin A$. The average power is

$$\frac{1}{m} \sum_{j \in A} \bar{\Phi} \left(\bar{\Phi}^{-1} \left(\frac{\alpha w_j}{m} \right) - \xi_j \right).$$

with constraint

$$\sum_{j \in A} w_j = m.$$

Choose \underline{w} to maximize

$$\pi = \frac{1}{m} \sum_{j \in A} \bar{\Phi} \left(\bar{\Phi}^{-1} \left(\frac{\alpha w_j}{m} \right) - \xi_j \right) - \lambda \left(m - \sum_{j \in A} w_j \right)$$

by setting the derivative to zero

$$\begin{aligned} \frac{\partial}{\partial w_i} \pi &= -\lambda + \frac{\phi \left(\bar{\Phi}^{-1} \left(\frac{\alpha w_j}{m} \right) - \xi_j \right)}{\phi \left(\bar{\Phi}^{-1} \left(\frac{\alpha w_j}{m} \right) \right)} \frac{\alpha}{m} = 0 \\ \frac{m\lambda}{\alpha} &= \frac{\phi \left(\bar{\Phi}^{-1} \left(\frac{\alpha w_j}{m} \right) - \xi_j \right)}{\phi \left(\bar{\Phi}^{-1} \left(\frac{\alpha w_j}{m} \right) \right)} \end{aligned}$$

The \underline{w} that solves these equations is given in (9). Finally, solve for c such that $\sum_i w_i = m$. ■

Proof of Lemma 3.3. Choose $K > 1$ such that $1/(K+1) < 1/a - \epsilon$. Choose $1 > \gamma > (2\alpha - a)/K$. Choose a small $c > 0$. Let $\xi = A + \sqrt{A^2 - 2c}$ and $u = B - \sqrt{B^2 - 2c}$ where

$$A = \bar{\Phi}^{-1} \left(\frac{\alpha}{(m(\gamma K + a))} \right), \quad B = \bar{\Phi}^{-1} \left(\frac{K\alpha}{(m(\gamma K + a))} \right). \quad (48)$$

Then $\rho(\xi) = 1/a$ and $\tilde{\rho}(\xi) = 1/(K+1)$. Now $d(Q, \tilde{Q}) = \gamma$. Taking K sufficiently large and γ sufficiently close to $(2\alpha - a)/K$ makes $\gamma < \delta$. ■

Proof of Theorem 3.5. The first statement follows easily by noting that the worst case corresponds to choosing weight B in the first term in $R(\xi)$ and choosing weight b in the second term in $R(\xi)$. The rest follows by Taylor expanding $R_{b,B}(\xi)$ around $b = 1$. ■

Proof of Lemma 3.6. With $b = 0$, $R_{b,B}(\xi) \geq 0$ when

$$\bar{\Phi}(z_{B\alpha/m} - \xi) - 2\bar{\Phi}(z_{\alpha/m} - \xi) \geq 0. \quad (49)$$

With $B \geq 2$, (49) holds at $\xi = 0$. The left hand side is increasing in ξ for ξ near 0 but (49) does not hold at $\xi = z_{\alpha/m}$. So (49) must hold in the interval $[0, \xi_*]$. Rewrite (49) as $\bar{\Phi}(z_{B\alpha/m} - \xi) - \bar{\Phi}(z_{\alpha/m} - \xi) \geq \bar{\Phi}(z_{\alpha/m} - \xi)$. We lower bound the left hand side and upper bound the right hand side.

The left hand side is $\overline{\Phi}(z_{B\alpha/m} - \xi) - \overline{\Phi}(z_{\alpha/m} - \xi) = \int_{z_{B\alpha/m} - \xi}^{z_{\alpha/m} - \xi} \phi(u) du \geq (z_{\alpha/m} - z_{B\alpha/m})\phi(z_{\alpha/m} - \xi)$. The right hand side can be bounded using Mill's ratio: $\overline{\Phi}(z_{\alpha/m} - \xi) \leq \phi(z_{\alpha/m} - \xi)/(z_{\alpha/m} - \xi)$. Set the lower bound greater than the upper bound to obtain the stated result. ■

It is convenient to prove Theorem 3.8 before proving Theorem 3.7.

Proof of Theorem 3.8. Let c_* solve

$$\gamma \overline{\Phi}(\sqrt{2c_*}) + a \overline{\Phi}\left(\frac{\xi}{2} + \frac{c_*}{\xi}\right) = \frac{\alpha}{m}. \quad (50)$$

We claim first that for any $c > c_*$, there is no u such that the weights average to 1. Fix $c > c_*$. The weights average to 1 if and only if

$$\gamma \overline{\Phi}\left(\frac{c}{u} + \frac{u}{2}\right) + a \overline{\Phi}\left(\frac{\xi}{2} + \frac{c}{\xi}\right) = \frac{\alpha}{m}. \quad (51)$$

Since $c > c_*$ and since the second term is decreasing in c , we must have

$$\overline{\Phi}\left(\frac{c}{u} + \frac{u}{2}\right) > \overline{\Phi}(\sqrt{2c_*}). \quad (52)$$

The function $r(u) = \overline{\Phi}(c/u + u/2)$ is maximized at $u = \sqrt{2c}$. So $r(\sqrt{2c}) \geq r(u)$. But $r(\sqrt{2c}) = \overline{\Phi}(\sqrt{2c})$. Hence $\overline{\Phi}(\sqrt{2c}) \geq r(u) \geq \overline{\Phi}(\sqrt{2c_*})$. This implies $c < c_*$ which is a contradiction. This establishes that $\sup_u c(u) \leq c_*$. On the other hand, taking $c = c_*$ and $u = \sqrt{2c_*}$ solves equation (51). Thus c_* is indeed the largest c that solves the equation which establishes the first claim. The second claim follows by noting that

$$\gamma \overline{\Phi}(\sqrt{2c_*}) + a \overline{\Phi}\left(\frac{\xi}{2} + \frac{c_*}{\xi}\right) = \gamma \overline{\Phi}(\sqrt{2c_*}) + O(a). \quad (53)$$

Now set this expression equal to α/m and solve. ■

Proof of Theorem 3.7. Define c_* as in (50). If $u_* = \sqrt{2c_*} \leq \xi$ then the the proof proceeds as in the previous proof. So we first need to establish for which values of ξ is this true. Let $r(c) = \gamma \overline{\Phi}(\sqrt{2c}) + a \overline{\Phi}(\xi/2 + c/\xi)$. We want to find out when the solution of $r(c) = \alpha/m$ is such that $\sqrt{2c} \leq \xi$, or equivalently, $c \leq \xi^2/2$. Now r is decreasing in c . Since $\gamma + a \geq \alpha/m$, $r(-\infty) \geq \alpha/m$. Hence there is a solution with $c \leq \xi^2/2$ if and only if $r(\xi^2/2) \leq \alpha/m$. But $r(\xi^2/2) = (\gamma + a)\overline{\Phi}(\xi)$ so we conclude that there is such a solution if and only if $(\gamma + a)\overline{\Phi}(\xi) \leq \alpha/m$, that is, $\xi \geq z_{\alpha/(m(\gamma+a))} = \xi_0$.

Now suppose that $\xi < \xi_0$. We need to find $u \leq \xi$ to make c as large as possible in the equation $v(u, c) \equiv \gamma \bar{\Phi}(u/2 + c/u) + a \bar{\Phi}(\xi/2 + c/\xi) = \alpha/m$. Let $u_* = \xi$ and $c_* = \xi z_{\alpha/(m(\gamma+a))} - \xi^2/2$. By direct substitution, $v(u_*, c_*) = \alpha/m$ for this choice of u and c and clearly $u_* \leq \xi$ as required. We claim that this is the largest possible c_* . To see this, note that $v(u, c) < v(u, c_*)$. For $\xi \leq \xi_0$, $v(u, c_*)$ is a decreasing function of u . Hence, $v(u, c) < v(u, c_*) \leq v(u_*, c_*) = \alpha/m$. This contradicts the fact that $v(u, c) = \alpha/m$.

For the second claim, note that the power of the weighted test beats the power of Bonferroni if and only if the weight $w = (m/\alpha) \bar{\Phi}(\xi/2 + C(\xi)/2) \geq 1$ which is equivalent to

$$C(\xi) \leq \xi z_{\alpha/m} - \xi^2/2. \quad (54)$$

When $\xi \leq \xi_0$, $C(\xi) = \xi \xi_0 - \xi^2/2$. By assumption, $\gamma + a \leq 1$ so that $z_{\alpha/(m(\gamma+a))} \leq z_{\alpha/m}$ and Now suppose that $\xi_0 < \xi \leq \xi_*$. Then $C(\xi)$ is the solution to $r(c) = \gamma \bar{\Phi}(\sqrt{2c}) + a \bar{\Phi}(\xi/2 + c/\xi) = \alpha/m$. We claim that (54) still holds. Suppose not. Then, since $r(c)$ is decreasing in c , $r(\xi z_{\alpha/m} - \xi^2/2) > r(C(\xi)) = \alpha/m$. But, by direct calculation, $r(\xi z_{\alpha/m} - \xi^2/2) > \alpha/m$ implies that $\xi > \xi_*$ which is a contradiction. Thus (28) holds.

Finally, we turn to (29). In this case, $C(\xi) = z_{\alpha/(m\gamma)}^2/2 + O(a)$. The worst case power is $\bar{\Phi}(C(\xi)/\xi - \xi/2) = \bar{\Phi}(z_{\alpha/(m\gamma)}^2/(2\xi) - \xi/2) + O(a)$. The latter is increasing in ξ and so is at least $\bar{\Phi}(z_{\alpha/(m\gamma)}^2/(2\xi_*) - \xi_*/2) + O(a) = \bar{\Phi}((z_{\alpha/(m\gamma)}^2/(2\xi_*) - \xi_*^2)/(2\xi_*)) + O(a)$ as claimed. The next two equations follow from standard tail approximations for Gaussians. Specifically, a Gaussian quantile $z_{\beta/m}$ can be written as $z_{\beta/m} = \sqrt{2 \log(m L_m / \beta)}$ where $L_m = c \log^a(m)$ for constants a and c (Donoho and Jin 2004). Inserting this into the previous expression yields the final expression. ■

Proof of Theorem 4.2. Setting $\pi(w, \xi_m) = \bar{\Phi}(\bar{\Phi}^{-1}(w\alpha/m) - \xi_m)$ equal to $1 - \beta$ implies $w = (m/\alpha) \bar{\Phi}(z_{1-\beta} + z_{\alpha/m})$ which is equal to w_1 as stated in the theorem. The stated form of w_0 implies that the weights average to 1. The stated solution thus satisfies the restriction that a fraction ϵ have power at least $1 - \beta$. Increasing the weight of any hypothesis whose weight is w_0 necessitates reducing the weight of another hypothesis. This either reduces the minimum power of forces a hypothesis with power $1 - \beta$ to fall below $1 - \beta$. Hence, the stated solution does in fact maximize the minimum power. ■

The proof of Theorem 4.3 is similar to the previous proof and is omitted.

Proof of Theorem 5.2. The familywise error is

$$\mathbb{P}(\mathcal{R} \cap \mathcal{H}_0) \leq \sum_{j \in \mathcal{H}_0} \mathbb{P} \left(P_j \left(T_j^{(1)}, T_j^{(2)} \right) \leq \frac{w_j(T_j^{(1)})\alpha}{m} \right) \quad (55)$$

$$= \sum_{j \in \mathcal{H}_0} \mathbb{E} \left(\mathbb{P} \left(P_j \left(t_j^{(1)}, T_j^{(2)} \right) \leq \frac{w_j(t_j^{(1)})\alpha}{m} \mid T_j^{(1)} = t_j^{(1)} \right) \right) \quad (56)$$

$$= \sum_{j \in \mathcal{H}_0} \mathbb{E} \left(\mathbb{P} \left(\bar{\Phi}(t_j^{(1)}, T_j^{(2)}) \leq \frac{w_j(t_j^{(1)})\alpha}{m} \mid T_j^{(1)} = t_j^{(1)} \right) \right) \quad (57)$$

$$= \sum_{j \in \mathcal{H}_0} \mathbb{E} \left(\mathbb{P} \left((1-b)^{1/2} T_j^{(2)} \geq \bar{\Phi}^{-1} \left(\frac{w_j(t_j^{(1)})\alpha}{m} \right) - b^{1/2} t_j^{(1)} \right) \right) \quad (58)$$

$$= \sum_{j \in \mathcal{H}_0} \mathbb{E} \left(\bar{\Phi} \left(\frac{\bar{\Phi}^{-1} \left(\frac{w_j(t_j^{(1)})\alpha}{m} \right) - b^{1/2} t_j^{(1)}}{(1-b)^{1/2}} \right) \right) \quad (59)$$

$$= \sum_{j \in \mathcal{H}_0} \mathbb{E} \left(\bar{\Phi} \left(\frac{\frac{\hat{\xi}_j}{2} + \frac{c(t^{(1)})}{\hat{\xi}_j} - b^{1/2} t_j^{(1)}}{(1-b)^{1/2}} \right) \right) \quad (60)$$

$$\leq \sum_{j=1}^m \mathbb{E} \left(\bar{\Phi} \left(\frac{\frac{\hat{\xi}_j}{2} + \frac{c(t^{(1)})}{\hat{\xi}_j} - b^{1/2} t_j^{(1)}}{(1-b)^{1/2}} \right) \right) \quad (61)$$

$$\leq \mathbb{E} \sum_{j=1}^m \left(\bar{\Phi} \left(\frac{\frac{\hat{\xi}_j}{2} + \frac{c(t^{(1)})}{\hat{\xi}_j} - b^{1/2} t_j^{(1)}}{(1-b)^{1/2}} \right) \right) \quad (62)$$

$$= \alpha. \quad \blacksquare \quad (63)$$

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.
- Benjamini, Y. and Hochberg, Y. (1997). Multiple Hypothesis Testing with Weights, *Scandinavian Journal of Statistics*, **24**, 407–418.
- Chen, James J. and Lin, Karl K. and Huque, Mohammad and Arani, Ramin B. (2000). Weighted p -value adjustments for animal carcinogenicity trend test. *Biometrics*, 56, 586–592.
- Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, **32**, 962–994.
- Genovese, C. R., Roeder, K. and Wasserman, L. (2005). False Discovery Control with P-Value Weighting. To appear: *Biometrika*.
- Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.
- Roeder, Bacanu, Wasserman and Devlin (2005). Using Linkage Genome Scans to Improve Power of Association in Genome Scans. *The American Journal of Human Genetics*. **78**.
- Rubin, D, van der Laan, M. and Dudoit, S. (2005). Multiple testing procedures which are optimal at a simple alternative. Technical report 171, Division of Biostatistics, School of Public Health, University of California, Berkeley .
- Skol A.D., Scott L.J., Abecasis G.R. and Boehnke M. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*. 38:390-394.
- Signoravitch, J. (2006). Optimal multiple testing under the general linear model. Technical report. Harvard Biostatistics.
- Storey J.D. (2005). The optimal discovery procedure: A new approach to simultaneous significance testing. UW Biostatistics Working Paper Series, Working Paper 259.
- Thomas D.C., Haile R.W. and Duggan D. (2005) Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet* 77:337-345